MULTI-FOCUS ULTRASOUND IMAGING USING GENERATIVE ADVERSARIAL NETWORKS

Sobhan Goudarzi, Amir Asif, Hassan Rivaz

Department of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada.

ABSTRACT

Ultrasound (US) beam can be focused at multiple locations to increase the lateral resolution of the resulting images. However, this improvement in resolution comes at the expense of a loss in frame rate, which is essential in many applications such as imaging moving anatomy. Herein, we propose a novel method based on generative adversarial network (GAN) for achieving multi-focus line-per-line US image without a reduction in the frame rate. Results on simulated phantoms as well as real phantom experiments show that the proposed deep learning framework is able to substantially improve the resolution without sacrificing the frame rate.

Index Terms— ultrasound imaging, focal point, frame rate, generative adversarial network, adversarial loss.

1. INTRODUCTION

There are three main US imaging approaches namely: (1) Classical focused transmission (also known as line-per-line acquisition); (2) Plane wave transmissions; and (3) Element-by-element transmission based synthetic aperture imaging. Focusing can be done through electronic beamforming or a lens in front of the aperture. In classical focused transmission, US wave has a complex bowtie shape with sidelobes and grating lobes. The US image formation is based on the assumption that received echoes stem from within the main transmitted US beam. In practice, a strong reflector from outside of the main lobe may generate detectable echoes resulting it being falsely displayed. Hence, narrower transmitted beams result in improved resolution and less artifacts.

In classical focused transmission, the quality of reconstructed image is optimal at the focal point but degrades progressively away from it. Consequently, to preserve the quality along the axial direction, several images with different transmit focal depths can be merged together. In this procedure, the frame rate proportionally decreases with an increase in the number of focal points.

Using a single transducer element for emission with low emitted energy, synthetic aperture imaging has a poor signal-to-noise ratio and a limited depth of penetration. Although it has been proved [1] that optimal multi-focus US images and a high frame rate can simultaneously be achieved by coherently compounding plane waves transmitted with different angels, clinical application of this method needs costly data acquisition boards, large data transfer bandwidth, and powerful parallel processing units.

Inspired by the success of deep learning, we propose a data-driven method for multi-focus line-per-line US imaging without a loss of frame rate. More specifically, we train a Generative Adversarial Network (GAN) [2] to learn propagation of US waves in the tissue.

Convolutional neural networks are able to extract necessary information (features) from raw data without engineering hand-drafted features. The main idea of GANs is surrendering the task of defining an objective function for the system. In particular, GANs consist of generator and discriminator networks, which compete with each other. The discriminator provides the generator with the quality of generated data, and the generator tries to fool discriminator by generating more realistic data [2]. Hence, the discriminator is the objective function for the generator while itself is also interestingly trained during a single training process. GAN has been successfully applied in different tasks such as denoising, super-resolution, and medical image synthesis [3, 4, 5, 6].

Herein, we propose a novel approach generating several focal points by sending a single focused beam. To this end, fully convolutional networks are used as generators estimating other focal points through different GANs. Since US images are not stationary along the axial direction, the nonlinear propagation equation of the US beam for having a narrow beam everywhere is estimated through different GANs . Experiments are performed using both simulated and real phantom data. We show that high quality multi-focus US images can be generated without sacrificing the frame rate.

2. MULTI-FOCUS ULTRASOUND IMAGING

2.1. Focusing

In classic line-per-line imaging, a set of excitation pulses with proper time delays are applied to crystals in order to focus at a specific axial depth (z_0) shown in Fig. 1. By considering variations of acoustic potentials along the axial direction, the

This research is funded by NSERC Discovery Grants RGPIN-2015-04136 and RGPIN-2017-06629.

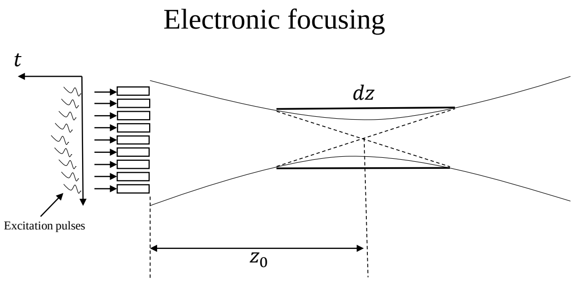

Fig. 1. Electronic focusing of an ultrasound beam.

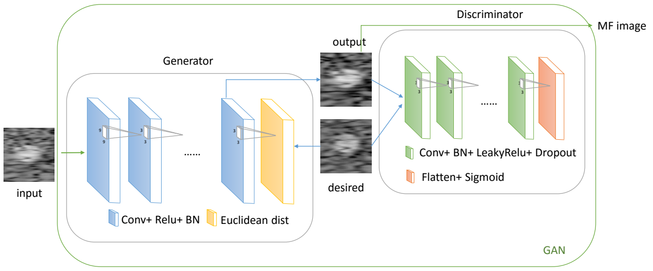

Fig. 2. The structure of the proposed GAN.

depth of focus (dz) can be defined as the distance between two points where the field on axis is 3dB less than at the focal point. In order to have an optimal multi-focus image (focused everywhere along the axial direction), the maximum distance between transmitted focal points has to be equal to the depth of focus. Therefore, we formulate our problem as finding a nonlinear function, which transforms the bowtie-shaped focused beam (with one focal point) to a thin cylindrical beam. As this nonlinear function is nonstationary along the axial direction, its parameters vary as a function of depth. As such, different networks corresponding to different depths need to be trained. Consequently, the method proposed here is based on partial estimation of a nonlinear function for multiple depth intervals, which is a common solution for addressing nonstationary problems. In other words, we break the image into a number of limited intervals along the axial direction such that the stationary assumption in training convolutional neural networks is valid, and subsequently train a GAN for each interval.

2.2. GENERATIVE ADVERSARIAL NETWORK

2.2.1. *Background*

GAN training is a min-max game wherein the generator estimates the input-output function and the discriminator distinguishes between real and synthesized data. The optimization is done in an alternating manner to solve the following adversarial objective function [2]:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}\left[log\ D(x)\right] + \mathbb{E}_{z\sim p_z(z)}\left[log(1-D(G(z))\right] \qquad (1)$$

where x and z are, respectively, the desired and input data. E denotes the expected value, and D and G are, respectively, the discriminator and generator. The generator tries to operate on the input in such a way that output is similar to the desired output such that D cannot discriminate it. Notation p(.) denotes probability of the enclosed parameter. In other words, the output resides in the manifold of real images and is classified as real. It has to be mentioned that we used non-saturating GAN to have stronger gradient for G and prevent saturation. More specifically, G is trained to maximize $log(D(G(z)))$ instead of training G to minimize $log(1-D(G(z))$. As a result, much stronger gradients are provided early in learning while the same fixed point of the dynamics of G and D are preserved [2].

2.2.2. *Proposed network*

Our proposed network is shown in Fig. 2. GANs are one of the most active areas of research in deep learning. Previous work has culminated in several guidelines to prevent mode collapse and non-convergence [7, 8], which are taken into account in our work. The generator in Fig. 2 is a fully convolutional network consisting of 9 layers, where the layers respectively contains 32, 32, 32, 64, 64, 64, 32, 32, and 1 filters with square kernels of size 9, 3, 3, 3, 9, 3, 3, 7, 3. Each layer also contains ReLU activation functions and a batchnorm layer. The discriminator in Fig. 2 consists of 4 layers containing 32, 64, 128, and 256 convolution filters with the same kernel size of 3. Each layer also contains LeakyReLU, batchnorm, and dropout (rate = 0.25) layers. The last layer is flattening with sigmoid activation for getting the output label. The number of filters and layers was chosen to maintain a minimum number of parameters for preserving the generalization performance and a more stable training. Kernel sizes were chosen empirically. We did not encounter checkerboard artifacts because the input and output patches have the same size.

2.2.3. *Training*

We first normalized the intensity input US images to [-1,1], and then after shuffling, a patch of input-target pairs of size 50×50 were extracted. In each iteration, the discriminator is trained before the generator. To prevent mode collapse [7], each patch is split into two parts. First, the discriminator is trained by the generator output and desired data, and then the generator is trained by both MSE and adversarial loss. The patch size is 20, and Adam optimizer with $\beta_1 = 0.5$ and learning rate of 10^{-4} is used. The code is implemented using TensorFlow library, and training was done with a Nvidia Titan Xp GPU.

The solution to training a GAN network (which is a game between two players) is a Nash equilibrium. In fact, by having the optimal discriminator, the global minimum of generator's

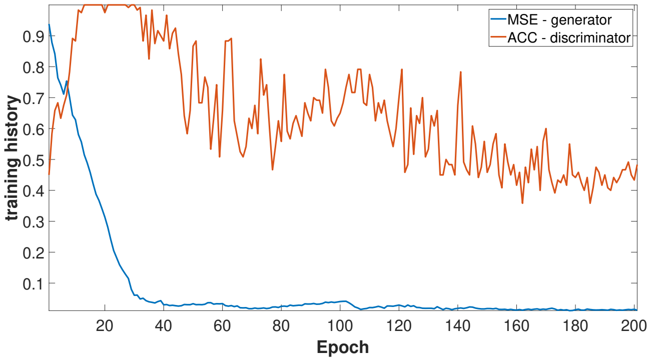

Fig. 3. Plot of the MSE loss of generator and accuracy of discriminator versus number of epochs during training.

loss function is achieved if and only if $p_g = p_{data}$, which means that the discriminator gives the same probability of 0.5 to both generated and real data. Although the two players may suddenly reach an equilibrium, the training process oscillates between two modes and players repeatedly undo each other. Here, we used three strategies for gaining the best training performance. First, we found that the learning rate of the adversarial loss function should be 10^{-3}, which prevents MSE to be the dominant objective. Second, the learning rate of all losses are reduced to 10% of corresponding initial values after 100 epochs in order to better probe the search space. Finally, we chose the interval of [0.49,0.51] for discriminator accuracy to stop the training process. Training history is illustrated in Fig. 3, which shows desirable oscillations in the discriminator and a relatively steady improvement in the generator.

3. EXPERIEMENTS

3.1. Evaluation setting

For evaluation, we place three real equispaced focal points in the axial direction of the US image, and blend the resulting three images by weighted spatial averaging as in commercial US machines. As such, the multi-focus image (desired) has 3 layers with 2 blended regions (Fig. 4 (b)). One of the images (Fig. 4 (a)) with the middle focal point is the input of our model. Therefore, the middle layer of output (Fig. 4 (c)) comes exactly from the input, and two other layers are estimated from related layers of input through two GANs. Each layer is broken into 50×50 patches and fed to the network. During the test phase, we do not break the image, and each layer is fed to the generator to prevent the blocking artifact. For quantitative analysis, peak signal to noise ratio (PSNR), normalized root mean square error (NRMSE), and structural similarity (SSIM) index are calculated between ground truth and both of the output of proposed network and input.

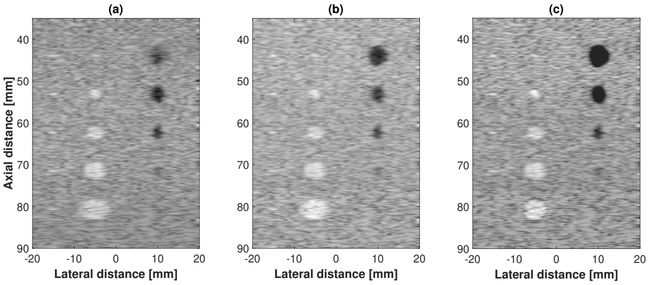

Fig. 4. Results of the proposed method on simulated phantom data. (a) Input image with a single focal point. (b) Desired image with 3 focal points. (c) Output of the proposed model.

3.2. Dataset

Our dataset contains two groups of simulated and real phantom US images. US images were simulated using the Field II program [9]. The simulated phantoms consisted of a collection of point targets, five cyst regions, and five hyperechoic regions. The total number of simulated phantoms was 750, which were obtained from independent realizations of the scatterers within the phantom. For each realization (i.e., each phantom), three different images were simulated by changing the location of the focal point.

Real phantom data was collected from Multi-Purpose Multi-Tissue Ultrasound Phantom (CIRS model 040GSE, Norfolk, VA) using an E-CUBE 12 Alpinion machine with L3-12H high density linear array and a centre frequency of 8.5 MHz. The sampling frequency of the radio-frequency (RF) data is 40MHz, and 384 RF lines were collected for each image. 20 images were collected at different locations of the phantom. At each location, three images with different focal points were collected, while the probe was held with a mechanical arm to prevent any probe movement during changing the transmit focus point. This ensured that images with different focal depths were collected at the same location.

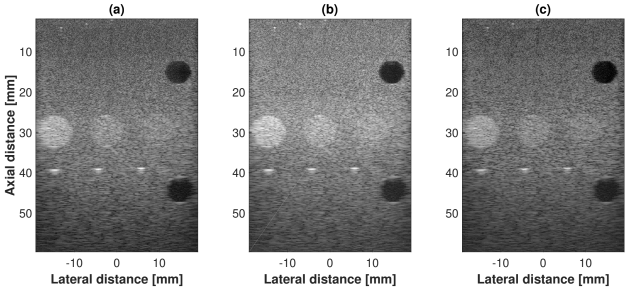

Fig. 5. Results of the proposed method on real phantom data. (a) Input image with a single focal point. (b) Desired image with 3 focal points. (c) Output of the proposed model.

Table 1. Performance indexes for input-desired and output-desired pairs.

data	input			estimated		
index	PSNR	NRMSE	SSIM	PSNR	NRMSE	SSIM
mean	24.34	0.0304	0.6157	25.53	0.0271	0.8098
std	0.79	0.0028	0.0212	1.934	0.0062	0.029
min	21.82	0.0223	0.5061	20.72	0.0144	0.7042
max	26.99	0.0405	0.6653	30.78	0.046	0.8684
median	24.3	0.0305	0.616	25.93	0.0252	0.8146

4. RESULTS

The entire database was broken into three sets of training, validation, and test groups with sizes of 70, 15, and 15 percent of the total size of images, respectively. The final model of training was saved and applied to the test set. For simulated images, Fig. 4 shows the output of the algorithm on a sample of the test set. As it can be seen, input, output, and desired images have the same quality in the middle axial region. However, the input image quality deteriorates in the shallow and deep regions, whereas the output and desired images have good quality throughout the images. It is clearly evident that the output of the proposed method even has better quality than desired around cyst regions and high scatterers. This improvement stems from our method of final model selection. In fact, as we never reach the perfect case (in which $p_g = p_{data}$), the model which has the best structural similarity to desired on validation dataset is chosen as final model among models within the accuracy interval of [0.49,0.51]. Table 1 summarizes the quantitative results. The PSNR, NRMSE, and particularly SSIM fully confirm the better quality of GAN output as was perceived in visual comparison of results (Fig. 4). It is worth mentioning that better perceptual quality was achieved through adversarial loss function. Fig. 3 shows that after epoch 40, there is no noticeable change in the MSE loss function and the discriminator mainly tunes the generator parameters to have a comparable output as desired.

For real phantom data, we used transfer learning because the number of images was limited. To this end, the final model of simulated data was used as the starting point of training on the training set of real phantom data, and the rest of precess is the same as before. Fig. 5 depicts the results on a sample test image of real phantom data. Fig. 5 shows the sharp boarders of cyst as well as the hyperechoic are preserved in the output of the model as the desired image. It has to be mentioned that in this experiment two first layers are estimated from the input and the last layer is the same as the input.

5. CONCLUSIONS

Increasing the number of focal points and breaking the US image to narrower axial layers is commonly used to preserve the depth of focus and lateral resolution throughout the image. This solution, however, substantially reduces the frame rate. In this paper, we proposed a novel GAN-based approach for having multi-focus US image in line-per-line imaging without a loss in frame rate. This approach can potentially be used in several applications that require both high resolution and high frame rate US imaging.

6. REFERENCES

[1] G. Montaldo, M. Tanter, J. Bercoff, N. Benech, and M. Fink, “Coherent plane-wave compounding for very high frame rate ultrasonography and transient elastography,” *IEEE Transactions on UFFC*, vol. 56, no. 3, pp. 489–506, March 2009.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, pp. 2672–2680. 2014.

[3] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, “Photo-realistic single image super-resolution using a generative adversarial network,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 105–114.

[4] D. Nie, R. Trullo, J. Lian, Li. Wang, C. Petitjean, S.Ruan, Q. Wang, and D. Shen, “Medical image synthesis with deep convolutional adversarial networks,” *IEEE Transactions on Biomedical Engineering*, 2018.

[5] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang, “Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss,” *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1348–1357, June 2018.

[6] O. Senouf, S. Vedula, G. Zurakhov, A. Bronstein, M. Zibulevsky, O. Michailovich, D. Adam, and D. Blondheim, “High frame-rate cardiac ultrasound imaging with deep learning,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 126–134.

[7] I. Goodfellow, “Nips 2016 tutorial: Generative adversarial networks,” *arXiv preprint arXiv:1701.00160*, 2016.

[8] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *arXiv preprint arXiv:1511.06434*, 2015.

[9] J. Jensen, “Field: A program for simulating ultrasound systems,” in *10th 10th Nordic-Baltic Conference on Biomedical Engineering*, 1996, vol. 4, pp. 351–353.